BAB II

TINJAUAN PUSTAKA

2.1 Penelitian Terdahulu

Dalam penelitian ini, peneliti menggunakan berbagai referensi studi sebelumnya untuk memperkuat landasan teoritis dan metodologis penelitian yang dilakukan. Upaya ini bertujuan untuk memastikan orisinalitas penelitian serta menghindari adanya duplikasi dengan penelitian yang sudah ada. Beberapa penelitian terdahulu yang relevan dijadikan acuan untuk mendukung dan memperkaya penelitian ini.

Penelitian yang dilakukan oleh Irma dkk., (2023) yang menggunakan algoritma *support vector machine* (SVM) untuk mengklasifikasikan ulasan pengguna pada aplikasi shopee. Hasil penelitian tersebut menunjukkan bahwa model SVM mampu memberikan performa yang tinggi dengan tingkat akurasi mencapai 98%.

Penelitian yang dilakukan oleh Rahmawati., dkk (2023) yang menggunakan algoritma *logistic regression* dengan studi kasus ulasan pengguna terhadap penerbangan lion air pada *platform* online. Hasil penelitian ini menunjukkan bahwa algoritma *logistic regression* memperoleh akurasi sebesar 82%. Meskipun algoritma ini terbukti efisien, peneliti juga mengemukakan bahwa ini rentan terhadap *underfitting* jika *dataset* yang digunakan tidak seimbang sehingga dapat mempengaruhi performa model.

Penelitian lain oleh Salsabilla., dkk (2022) yang menerapkan dua algoritma klasifikasi, yaitu SVM dan *naïve bayes* dengan studi kasus analisis sentimen tokoh Gus Dur. Pada penelitian ini digunakan metode SEMMA sebagai pendekatan analisis data. Hasil penelitian menunjukkan bahwa algoritma SVM memiliki performa yang lebih baik dibandingkan dengan *Naive Bayes*, dengan akurasi sebesar 84,27% untuk SVM dan 78,36% untuk *Naive Bayes*.

Selanjutnya, penelitian oleh Shufa dkk., (2020) yang membahas peningkatan performa algoritma *Naive Bayes* dalam klasifikasi sentimen terhadap

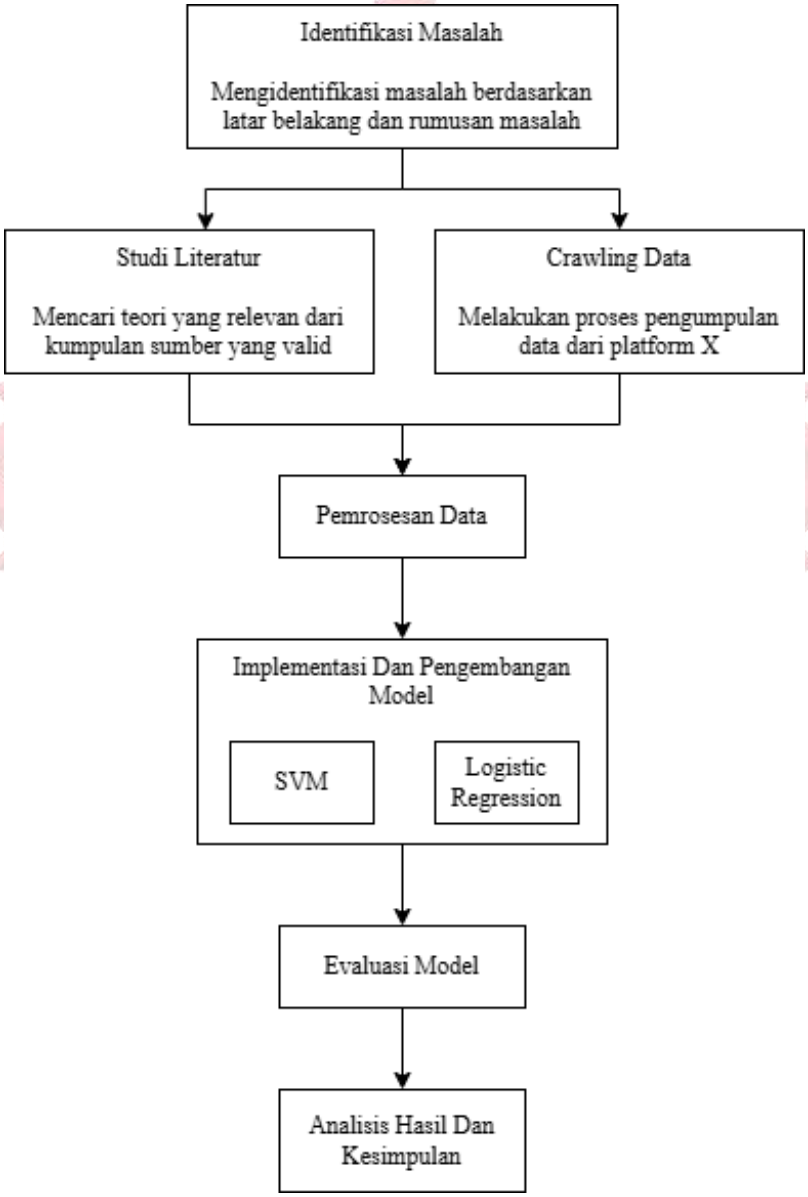
ketua umum partai politik di Indonesia. Penelitian ini menggunakan metode *adaptive boosting* untuk mengatasi masalah ketidakseimbangan data. Hasilnya menunjukkan adanya peningkatan akurasi dari 74,61% menjadi 79,17%, dengan rata-rata peningkatan performa sistem sebesar 8,70% setelah penerapan *adaptive boosting*.

Terakhir, penelitian oleh Nugraha., dkk (2019) meneliti tentang analisis sentimen terhadap pilpres 2019 dengan menggunakan algoritma *naïve bayes*. Hasil penelitian ini menunjukkan bahwa algoritma tersebut mendapatkan akurasi sebesar 56,67% dalam mengklasifikasi *tweet* dengan kata kunci "Jokowi Pilpres" dan akurasi sebesar 68,33% dalam mengklasifikasi tweet dengan kata kunci "Prabowo Pilpres".

Berdasarkan beberapa penelitian terdahulu yang telah dianalisis, terlihat bahwa analisis sentimen di media sosial telah dilakukan dengan beragam pendekatan algoritma seperti *Support Vector Machine* (SVM), *Logistic Regression*, dan *Naïve Bayes*. Namun demikian, penelitian ini memiliki sejumlah perbedaan yang menjadi ciri khas sekaligus kontribusi terhadap pengembangan studi sejenis, yaitu fokus pada waktu pada masa akhir jabatan Jokowi sebagai presiden Republik indonesia ke-7 dan proses pengumpulan data yang menggunakan metode *crawling* tanpa API resmi dari platform X, melainkan menggunakan library *tweet-harvest* yang memberikan fleksibilitas dan efisiensi dalam pengumpulan data.

2.2 Kerangka Pemikiran

Kerangka Pemikiran yang dimaksudkan dalam penelitian ini adalah tahapan-tahapan yang dilakukan untuk menganalisis dan mengklasifikasikan sentimen publik terhadap tokoh Jokowi dengan studi komparatif antara algoritma *Support Vector Machine* dan *Logistic Regression* dalam proses tahapannya. Berikut alur penelitian ini digambarkan pada Gambar 2.1



Gambar 2.1 Kerangka Pemikiran

1) Indentifikasi Masalah

   Pada tahap ini menetapkan masalah utama dalam penelitian ini berdasarkan latar belakang dan rumusan masalah.

2) Studi Literatur

   Pada tahap ini melakukan penelusuran terhadap penelitian sebelumnya yang relevan ataupun sumber-sumber yang valid guna mendapatkan landasan teori yang sesuai dengan masalah yang diteliti.

3) Pengumpulan Data

   Pada tahap ini dilakukan pengumpulan data dari sosial media X menggunakan teknik *crawling*.

4) Pemrosesan Data

   Pada tahap ini akan dilakukan pemrosesan data mentah menjadi data tersetruktur agar siap digunakan dalam pembuatan model.

5) Implementasi dan Pengembangan Model

   Pada tahap ini akan dilakukan pembuatan dan pengembangan model menggunakan algoritma SVM dan *Logistic Regression*.

6) Evaluasi Model

   Pada tahap ini dilakukan evaluasi model menggunakan matrik evaluasi seperti akurasi, presisi, *recall* dan *F-1 score*. Selanjutnya akan dibandingkan hasil performa antara model yang dibuat.

7) Analisis Hasil dan Kesimpulan

   Pada tahap ini akan dilakukan analisis terhadap hasil evaluasi model dan akan diambil kesimpulan.

2.3 Landasan Teori

2.3.1 Analisis Sentimen

Analisis sentimen adalah salah satu contoh dari bidang *Natural Language Processing (NLP)*. *Natural Language Processing (NLP)* merupakan bidang ilmiah yang membahas tentang cara kerja komputer agar bisa berfikir layaknya manusia dan merupakan bagian dari *Artifical Intellegence (AI)* atau kecerdasan buatan. *Artifical Intellegence* merupakan salah satu cabang dari *data mining* dan dalam

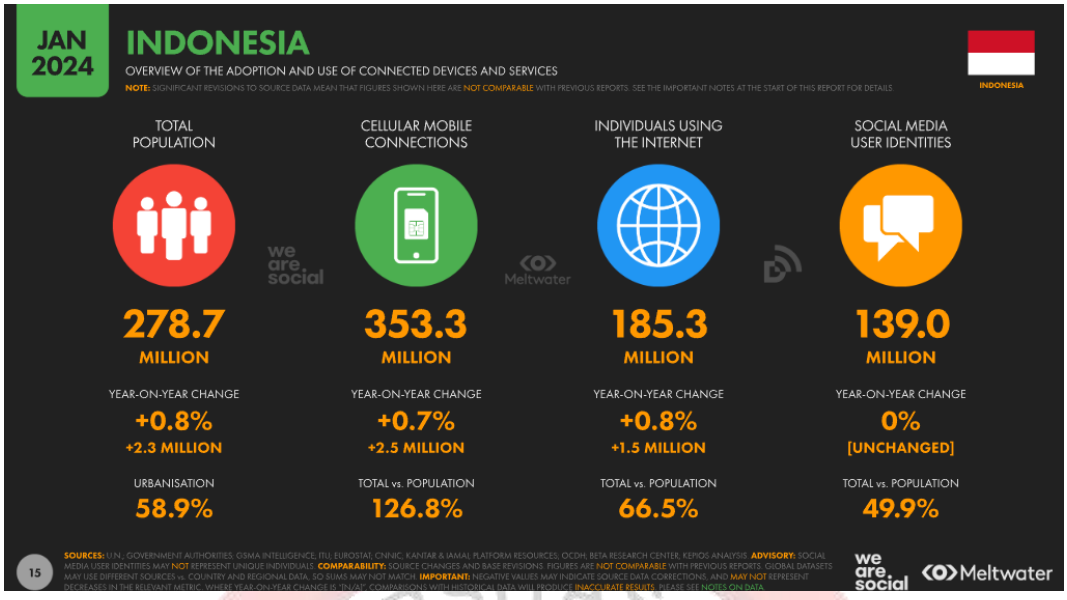
penerapannya memerlukan *machine learning*. *Machine learning* dapat digunakan untuk mengambil keputusan menggantikan manusia karena *machine learning* tidak mempunyai perasaan seperti manusia sehingga keputusan yang diambil berdasarkan dari data yang sudah diolah (Saputra dkk, 2022).

Analisis sentimen merupakan cabang dari penelitian *text mining* yang melakukan proses pengklasifikasian data teks. Analisis sentimen dapat melakukan ekstraksi pendapat, emosi dan evaluasi tertulis dari seseorang tentang topik tertentu menggunakan teknik pemrosesan bahasa alami (Irma dkk, 2023). Setiap dataset yang digunakan memerlukan penanganan yang berbeda-beda dalam analisis sentimen. Analisis sentimen ditujukan bukan hanya untuk pribadi, tetapi juga digunakan untuk mengatasi berbagai banyak hal seperti mengenai bisnis, program, produk, aplikasi, dan lain-lain yang dapat dikomentari publik (Herlinawati dkk., 2020). Salah satu keunggulan analisis sentimen adalah dapat menghemat waktu dan tenaga dalam melakukan penelitian dengan jumlah data yang besar (Nur Adinda, 2022).

Secara umum, analisis sentimen terbagi menjadi lima langkah yaitu *crawling data, pre-processing data, feature selection, classification dan evaluation*. Manfaat adanya analisis sentimen yaitu sebagai evaluasi dan ide pada berbagai bidang (Nur Adinda, 2022). Hasil dari analisis sentimen juga dapat menjadi sebuah gambaran bagi perusahaan, *public figure*, dan pemerintahan untuk menentukan langkah selanjutnya (Natasuwarna, 2020).

2.3.2 Media Sosial X (twitter)

Masyarakat dapat mengungkapkan sentimen atau opini terhadap suatu fenomena yang terjadi melalui berbagai media baik media konvensional maupun media digital. Salah satu media yang sering digunakan adalah media sosial (Shufa dkk., 2020). Dalam laporan " *Digital Around The World* 2024" tertuang bahwa dari total 278,7 juta penduduk di Indonesia, sebanyak 139 juta orang diantaranya adalah pengguna media sosial, sehingga angka pengguna sebanyak 49,9% dari total jumlah penduduk di Indonesia yang disajikan pada Gambar 2.2.



Gambar 2.2 Data Pengguna Media Sosial di Indonesia

(sumber: www.datareportal.com)

Berbagai media sosial yang digunakan penduduk dunia diantaranya yaitu Facebook, Instagram, X, Threads dan Reddit. Dari semua media sosial tersebut, X menjadi salah satu media sosial yang sering digunakan untuk menyampaikan opini dan berdiskusi (Rusdiaman dkk, 2019). Sosial media X mempermudah masyarakat untuk bebas berpendapat dalam topik apapun melalui cuitan atau yang biasa disebut dengan *tweet* (Widowati dkk, 2020). Istilah *tweet* yaitu pengguna X dapat memberikan kabar terbaru, berekspresi, beraspirasi, dan beropini yang ditulis oleh pengguna X lainnya (Sari dkk,2019). X mempunyai kelebihan yaitu jangkauan yang luas, dapat menjangkau *publik figure*, media promosi lebih luas, banyak jaringan, dan lebih mudah diukur kemampuannya (Nur Adinda, 2022).

Berdasarkan informasi dari situs resmi X (twitter), X merupakan layanan bagi teman, keluarga, dan teman sekerja untuk berkomunikasi dan tetap terhubung melalui pertukaran pesan yang cepat. Pengguna memposting *tweet* (kicauan), yang dapat berisi foto, video, tautan, dan teks (Shufa dkk., 2020). Berikut fitur atau istilah – istilah yang ada pada sosial media X (Taufik dkk., 2018):

1) Trending topik adalah fitur yang menampilkan topik teratas berupa *hastag* yang sedang hangat dan banyak dibicarakan oleh pengguna X.

2) *Hastag* adalah fitur yang mengelompokkan *tweet* atau pesan.
3) *Following* adalah fitur pertemanan untuk saling mengikuti yang menghubungkan antara pengguna X.
4) *Retweet* adalah fitur untuk membagikan sebuah *twee*t pengguna lain
5) *Mention* adalah fitur untuk menyebut pengguna lain dalam sebuah postingan *tweet.*

2.3.3 Tokoh Jokowi

Jokowi atau yang mempunyai nama lengkap Joko Widodo adalah tokoh politik yang mempunyai pengaruh besar dalam dunia politik di Indonesia. Menurut situs resmi presiden RI, Jokowi lahir di Surakarta, Jawa Tengah pada tanggal 21 Juni 1961. Beliau menempuh studi pada fakultas kehutanan di universitas gadjah mada dan berhasil lulus di tahun 1985. Beliau memiliki istri bernama Iriana dan dari hasil pernikahannya dikaruniai 3 orang anak yaitu, Gibran Rakabuming Raka, Kahiyang Ayu, dan Kaesang Pangarep.

Sebelum terjun ke dunia politik, Jokowi berprofesi sebagai tukang kayu yang bergerak pada bidang usaha mabel. Karir politiknya dimulai sebagai Wali Kota Surakarta di tahun 2005–2012. Pamor Jokowi melambung tinggi setelah dirinya terpilih menjadi Gubernur DKI Jakarta di tahun 2012–2014. Pada tahun 2014, ia terpilih sebagai Presiden RI bersama Jusuf Kalla sebagai Wakil Presiden, dan pada 2019 terpilih kembali dengan Ma'ruf Amin sebagai Wakil Presiden yang menjadikkannya sebagai presiden RI ke-7. Perjalanan beliau dalam dunia politik tak luput dari perhatian publik dengan beragam opini dan argumen yang muncul (Seran dkk., 2024). Lembaga survei nasional juga telah melakukan peninjauan pada masyarakat Indonesia dan menemukan adanya pro kontra terhadap Joko Widodo (Seran dkk., 2024).

2.3.4 *Machine Learning*

Machine learning merupakan cabang dari kecerdasan buatan (*Artificial Intelligence*) yang memungkinkan komputer untuk belajar dari data dan membuat keputusan tanpa perlu pemograman secara eksplisit(Wardhana dkk., 2023). Model

machine learning belajar langsung melalui data dari data proses pelatihan yang memungkinkan sistem dapat menyesuaikan terhadap pola dan perubahan data yang bersifat dinamis. *Machine learning* dapat digunakan dalam berbagai bidang, seperti prediksi, klasifikasi dan pengenalan pola karena efisiensi proses dalam menganalisis data (Wardhana dkk., 2023).

2.3.5 Metode SEMMA

Metode SEMMA berfokus pada modifikasi, penambangan data, dan pemodelan yang dirancang untuk membantu pengguna *software* SAS *enterprise miner* (Komang dkk., 2022). Tahapan metode SEMMA adalah sebagai berikut (Dwison dkk., 2020) :

a. *Sample*

   Tahapan ini dilakukan dengan mencari teori-teori terkait penelitian ini yang bersumber dari jurnal, buku, ataupun situs-situs yang berkaitan. Data didapatkan dengan *crawling* data pada media sosial twitter.

b. *Explore*

   Tahap ini menjelaskan deskripsi data dan visualisasi data. Deskripsi data merupakan penjelasan dari gambaran data informasi yang akan digunakan.

c. *Modify*

   Pada tahap ini berupa modifikasi data dengan cara memilih, membuat, dan melakukan transformasi terhadap data yang diolah.

d. *Model*

   Pada tahap ini melakukan pembuatan dan pelatihan model yang akan digunakan.

e. *Asses*

   Pada tahap ini mengevaluasi dari model yang telah dibuat dengan matrik evaluasi.

2.3.6 *Text Preprocessing*

Text Preprocessing merupakan proses pengolahan *dataset* sebelum data tersebut siap digunakan. Pada kenyataannya, kesalahan sistem saat pencatatan menjadikan data tidak bersih dan tidak beraturan seperti duplikasi data, tipe data

yang berbeda dan lain sebagainya. Oleh karena itu, data mentah harus diproses dan diolah agar data siap digunakan untuk proses selanjutnya. Semakin bersih data yang diproses, maka kemungkinan besar hasil data tersebut semakin akurat (Saputra dkk., 2022).

a) *Cleaning*

Cleaning merupakan tahapan awal pembersihan data seperti pengecekan duplikasi data, menyeleksi kata atau atribut dan lain-lain yang dapat mempengaruhi sentimen.

b) *Tokenize*

Tokenisasi merupakan pemecahan kata dalam sebuah kalimat pada dataset untuk mempermudah tahapan selanjutnya.

c) *Casefolding*

Casefolding merupakan tahap mengubah semua teks menjadi huruf kecil.

d) *Stopwords* Removal

Stopwords removal merupakan tahap menghapus kata-kata yang tidak memberikan makna signifikan seperti “dan” atau “di”

e) *Stemming*

Stemming merupakan tahap merubah kata menjadi bentuk dasar seperti “berjalan” menjadi “jalan”.

f) Pelabelan

Pelabelan merupakan tahap pembagian data berdasar kelas.

g) Vektorisasi

Vektorisasi merupakan tahap mengubah data teks menjadi numerik agar dapat diproses oleh algoritma *machine learning*.

2.3.7 *Support Vector Machine*

Support vector machine (SVM) merupakan algoritma *supervised learning* yang membutuhkan sampel data dan biasanya digunakan untuk tugas klasifikasi atau regresi. *Support vector machine (SVM)* merupakan algoritma yang dikembangkan oleh Boser, Guyon, dan Vapnik pada tahun 1992. SVM memiliki

konsep kombinasi dari teori komputasi sebelumnya. Menurut Hilda dkk., (2019) algoritma ini mengubah data latih ke dimensi yang lebih tinggi menggunakan pola non-linear. Algoritma ini mempunyai tingkat akurasi yang lebih tinggi jika dibandingkan dengan algoritma lain(Pratiwi dkk., 2021). SVM dapat bekerja pada dataset yang mempunyai dimensi tinggi menggunakan kernel trik (Pane dkk., 2021). Pada pola hasil pelatihan, metode ini meliputi *machine learning* berdasarkan *Structural Risk Minimization* (SRM). Menurut Oktavia dkk., (2023) *Support Vector Machine (SVM)* merupakan salah satu teknik dalam *machine learning* yang berdasarkan teori struktural pembelajaran. Algoritma ini bekerja dengan menemukan *hyperplane* terbaik untuk memisahkan kelas-kelas data. *Hyperplane* adalah bidang pemisah antara satu kelas dengan kelas lainnya (Nur Adinda, 2022). *Margin* adalah jarak antara *Support Vector Machine* dengan *hyperplan* (Santoso, 2021). Dalam hal ini berperan untuk memisahkan *tweet* positif dan *tweet* negatif. Simbol dan representasi *hyperplane* dalam algoritma *Support Vector Machine (SVM)* adalah sebagai berikut :

$$w \cdot x + b = 0$$

Penjelasan Simbol :

w : Vektor bobot atau normal terhadap *hyperplane*.

x : Vektor data input.

b : bias atau *intercept* yang menggeser posisi *hyperplane* dari titik asal.

= 0 : Menandakan persamaan ini adalah batas atau pemisah antar kelas.

Hyperplane dan *Margin*

Hyperplane Positif : $w \cdot x + b = 1$

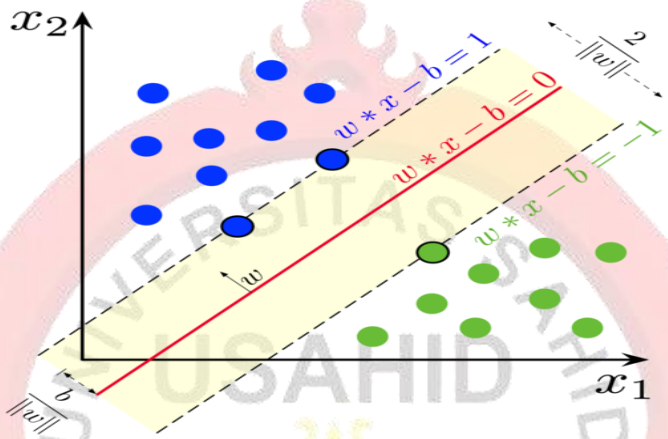
Hyperplane Negatif : $w \cdot x + b = 0$

Margin : Jarak antara dua *hyperplane* yang mendukung, yaitu:

$$Margin = \frac{1}{||w||}$$

||w|| adalah norma dari vektor w

Untuk memperjelas konsep tersebut, Gambar 2.3 berikut menampilkan ilustrasi visual mengenai *hyperplane*, *margin* dan distribusi data dua kelas yang umum digunakan dalam algoritma SVM. Hyperplane utama ditunjukkan dengan garis berwarna merah, sedangkan dua garis sejajar lainnya berfungsi sebagai batas margin. Titik-titik data yang menyentuh margin disebut *support vectors* karena keberadaannya menentukan posisi *hyperplane*.



Gambar 2.3 Ilustrasi Algoritma SVM

(Sumber: https://en.wikipedia.org/wiki/Support_vector_machine)

2.3.8 *Logistic Regression*

Logistic regression merupakan algoritma *supervised learning* yang digunakan untuk klasifikasi biner dengan memprediksi probabilitas suatu kelas berdasarkan fungsi logistik (*sigmoid*). Algoritma ini pertama kali dikembangkan dari konsep analisis regresi dan diperluas penggunaannya untuk masalah klasifikasi dengan mengestimasi kemungkinan terjadinya suatu hasil berdasarkan variabel input. *Logistic regression* merupakan sebuah algoritma yang bekerja berdasarkan hubungan antara satu variabel atau lebih (Novitasari dkk., 2019). Variabel *respons* dapat berupa kategori atau kualitatif, sedangkan variabel prediktornya dapat berupa kualitatif dan kuantitatif . Persamaan dalam *logistic regression* sebagai berikut :

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}$$

dimana :

$P(Y = 1|X)$ adalah probabilitas variabel dependen bernilai 1.

β0 adalah *intercept*.

β1 adalah koefisien regresi.

X adalah variabel independen.

Rumus *logistic regression* sebagai berikut :

$$\text{Accuracy} = \sum_{i=1}^{m} \frac{\left(y_{pred}^{(i)} == y_{true}^{(i)}\right)}{m}$$

dimana :

Accuracy: Mengukur proporsi prediksi yang benar terhadap seluruh prediksi. Nilai akurasi berkisar antara 0 hingga 1, atau dinyatakan dalam persen (0-100%).

∑i=1m: Menunjukkan bahwa perhitungan dilakukan dengan menjumlahkan seluruh data, dari i = 1 hingga i = m, di mana m adalah jumlah total data.

ypred(i): Label yang diprediksi oleh model untuk data ke-i.

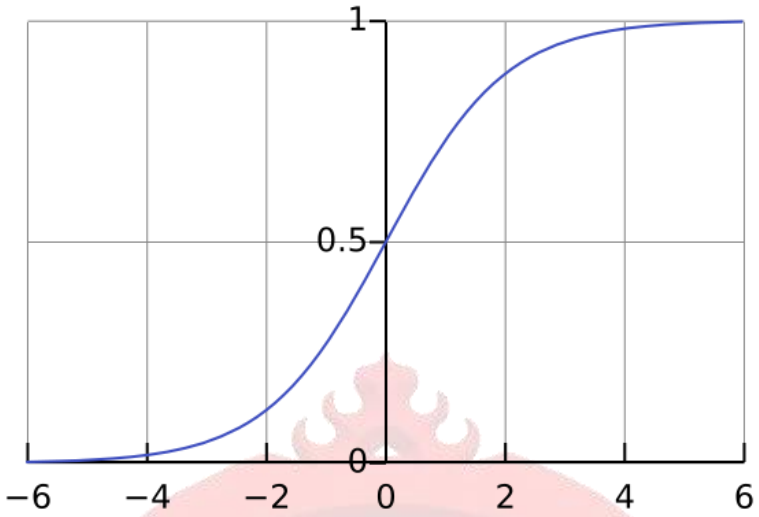
*ytrue(*i): Label sebenarnya (*ground truth*) untuk data ke-i.

Ekspresi (y_pred^(i) == y_true^(i)):

- Bernilai 1 jika y_pred^(i) = y_true^(i) (prediksi benar).
- Bernilai 0 jika y_pred^(i) ≠ y_true^(i) (prediksi salah).

m : Total jumlah sampel dalam *dataset*.

Untuk memperjelas fungsi logistik bekerja dalam proses klasifikasi, Gambar 2.4 berikut menampilkan bentuk kurva *sigmoid* yang digunakan dalam algoritma *Logistic Regression.*



Gambar 2.4 Ilustrasi Kurva *Sigmoid Logistic Regression*
(Sumber : https://commons.wikimedia.org/wiki/File:Logistic-curve.svg)

Kurva diatas menunjukkan skor *linear* dari kombinasi variabel independen dikonversi menjadi nilai probabilitas antar 0 hingga 1. Ilustrasi ini menggambarkan inti dari proses klasifikasi *biner* pada *Logistic Regression*, dimana fungsi *sigmoid* digunakan untuk memetakan semua nilai input kedalam rentang probabilistik yang dapat ditafsirkan sebagai keyakinan model terhadap suatu kelas.

2.3.9 *Confusion Matrix*

Confusion matrix adalah matriks yang digunakan untuk melakukan evaluasi proses model klasifikasi berupa jumlah data uji yang benar dan salah. Dengan adanya matriks ini dapat mengetahui kualitas kinerja model klasifikasi (Normawati dkk., 2021). *Confusion matrix* berisi berbagai performa yang dapat diukur seperti akurasi, presisi, *recall*, dan *F1 Score* untuk mengetahui kinerja dari pemodelan yang telah dilakukan sebelumnya (Saputra & Kristiyanti, 2022). Contoh *confusion matrix* menurut Normawati dkk., (2021) disajikan dalam Tabel 2.1 sebagai berikut:

Tabel 2.1 *Confusion Matrix*

Nilai Prediksi	Nilai Aktual	
	Positive	*Negative*
Positive	*True Positive*	*False Negatif*
Negative	*True Negative*	*False Negatif*

Tabel diatas merupakan tabel *confusion matrix* dengan penjelasan sebagai berikut:

1. *TP (True Positive)* = jumlah data nilai aktual kelas positif dan nilai prediksi kelas positif
2. *TN (True Negative)* = jumlah data nilai aktual negatif dan nilai prediksi negatif
3. *FP (False Positive)* = jumlah data nilai aktual positif dan nilai prediksi negatif
4. *FN (False Negative)* = jumlah data nilai aktual negatif dan nilai prediksi positif

2.3.10 *Lexicon Based*

Metode *lexicon-based* merupakan pendekatan yang menggunakan kamus (*lexicon*) yang telah diberi label, seperti positif, negatif, atau netral. Dalam konteks penilaian sentimen, dilakukan dengan menghitung jumlah kata positif dan negatif dalam suatu teks, kemudian menentukan polaritas berdasarkan dominasi kata-kata tersebut (Fauziah dkk., 2021).

2.3.11 *Grid Search*

Grid search merupakan teknik pencarian hyperparameter secara sistematis yang digunakan untuk mementukan kombinasi paramater terbaik pada *machine learning*. Menurut Raschka dkk., (2018) , metode ini berkerja dengan membuat "*grid*" dari semua kemungkinan nilai parameter, lalu mengevaluasi model menggunakan *cross validation* di setiap kombinasi. *Grid search* cocok digunakan untuk tuning parameter pada model yang tidak memiliki banyak *hyperparamete*r atau ketika *resource* komputasi terbatas (Bergstra dkk., 2012).

2.3.12 Google Colab *Research*

Google Colab adalah layanan *cloud computing* yang disediakan oleh Google untuk mendukung pengembangan dan penelitian ilmiah (Guntara, 2023). Colab memungkinkan untuk menulis dan mengeksekusi kode python arbitrer melalui *browser*, dan sangat cocok digunakan untuk *machine learning*, analisis data, dan pendidikan (Febby dkk., 2024).